Conservative bounds for the *pfd* of a 1-out-of-2 software-based system based on an assessor's subjective probability of "not worse than independence"

Bev Littlewood, Andrey Povyakalo

Centre for Software Reliability, City University, London EC1V 0HB

Abstract

We consider the problem of assessing the reliability of a 1-out-of-2 software-based system, in which failures of the two channels cannot be assumed to be independent with certainty. An informal approach to this problem assesses the channel *pfd*s (probabilities of failure on demand) *conservatively* and then multiplies these together in the hope that the conservatism will be sufficient to overcome any possible dependence between the channel failures. Our intention here is to place this kind of reasoning on a formal footing. We introduce a notion of "not worse than independence" and assume that an assessor has a prior belief about this, expressed as a probability. We obtain a conservative prior system *pfd*, and show how a conservative *posterior* system *pfd* can be obtained following the observation of a number of demands without system failure. We present some illustrative numerical examples, discuss some of the difficulties involved in this way of reasoning, and suggest some avenues of future research.

1 Background

We consider the problem of assessing the reliability of a 1-out-of-2 system in which the two software-based channels are "diverse" as a result of having been developed independently of one another (indeed, their designs may have been forced to be diverse by imposing diverse development procedures upon their designers). Such design-diverse fault tolerant systems have been used successfully in some safety critical applications: see (Littlewood, Popov et al. 2002; Wood, Belles et al. 2010).

Whilst there is some general evidence that this kind of design-diverse fault tolerance is a good way of achieving high reliability – for example from experiments – there are serious difficulties is assessing the reliability of a particular system. An important problem arises from the fact that we can never be certain that the channels in such a system will fail independently: so for a 1-out-of-2 system we cannot simply multiply together the channel *pfd*s to obtain the system *pfd*. In several experiments – see e.g. (Knight and Leveson 1986; Eckhardt, Caglayan et al. 1991) – "independently" developed software versions (channels) were shown to fail dependently. In fact there was a tendency for the dependence to be positive, i.e. the versions failed together more frequently than would have been the case if failures were independent. Even in these experiments, however, there was on average *some* benefit gained from the use of multiple channels (compared with single versions) (Knight and Leveson 1986), even if this was not as great as it would have been under independence.

The experimental results were confirmed in some contemporary theoretical modeling, which also provided a conceptual framework for understanding reasons for failure dependence (Eckhardt and Lee 1985; Littlewood and Miller 1989). The basic idea

introduced by Eckhardt and Lee is that "problem difficulty" varies over the demand space: some demands are "intrinsically harder" than others. That is, it is harder to build a program that executes such a demand correctly (i.e. the chance of a particular program doing so is smaller). If channel *A* fails on a randomly selected demand, one should conclude that this was probably a difficult demand and thus the chance of channel *B* failing on the same demand is greater than it otherwise would be: i.e. this *conditional* probability of *B* failing is great than *B*'s marginal *pfd*. The result is that there is positive association between channel failures, and the 1-out-of-2 system *pfd* is greater than it would be if independence of failures could be assumed.

Littlewood and Miller generalize this result to the case where diversity is forced by employing deliberately different "methodologies" to develop *A* and *B*. In this case the variation of difficulty for *A* will generally be different from that of *B*: demands that are hard for *B* may be easier for *A* and vice versa. It is shown that in this case the association between channel failures can be either positive or negative – i.e. it is possible to do *better* than the case of independence (the 1-out-of-2 system *pfd* can be smaller than the product of the two channel *pfd*s). Whether it is practically feasible to force the methodologies to be sufficiently different that the channels exhibit such negatively associated failure behaviour remains a moot point. If it is possible, it is unlikely that one could be *certain* that negative association of failures had been achieved for a particular pair of channels.

In summary, then, the position is this. Whilst there is evidence that this approach may be effective – in some average sense – in achieving system reliability, it is difficult to assess the reliability of a particular design-diverse system. This is because the level of association between the failures of the diverse channels will not be known – in particular it cannot be assumed that they will fail independently. These problems of assessment are important because they are a barrier to the use of what is otherwise one of the most promising approaches to very high system reliability.[1]

An interesting way around this difficulty arose in some discussions the authors had with engineers involved in the licensing of a 2-channel, 1-out-of-2 protection system. The *pfd* of the system was required to be no worse than 10^{-6}. It was expected that there would be extensive analysis of the "diversity-seeking" decisions involved in the designs of the two channels, so it may be reasonable to conclude that any dependence between the channel failure processes would be modest. The *pfd* claims for the two channels – 10^{-4} and 10^{-2} – were believed to be very conservative, sufficiently so that taking the product of these, it was claimed, would give a conservative value for the system *pfd* even in the possible presence of some positive dependence between the channels.

The difficulty with this kind of reasoning, we think, is that it makes a trade-off between very different things: conservatism in channel claims against optimism in claims about joint failure behaviour. It seems reasonable to ask *how* optimistic the independence claim is (i.e. how dependent the channel failures actually are) and *how*

---

[1] Although it should be said that other approaches to achieving high reliability also pose great difficulties in *assessing* what has been achieved in a particular instance. In fact, many claims for the efficacy of software engineering processes concern their "on average" performance, and what is achieved in a particular instance can be very different from this average. Also, it must be admitted that the empirical support, even for these average effects, is often weak.

pessimistic the channel claims are – and then to ask whether the latter is sufficient to overcome the former.

In the work reported here we aim to put this kind of reasoning on a more rigorous footing in order to make conservative claims for multi-channel systems in the presence of likely channel failure dependency. We begin with a brief examination of the nature of "association", or dependency, between channel failures, with the intent of explicitly modeling the uncertainty here.

Figure 1 illustrates the spectrum of possible dependence between channel failures. It ranges from a best case where there are no coincident failures (the failure regions of the input space for channel *A* and channel *B* are disjoint), to a worst case where all failures are coincident (the failure regions are identical). It is clear from the figure that independence is a very special case: it is just one point in the "middle" of this spectrum.

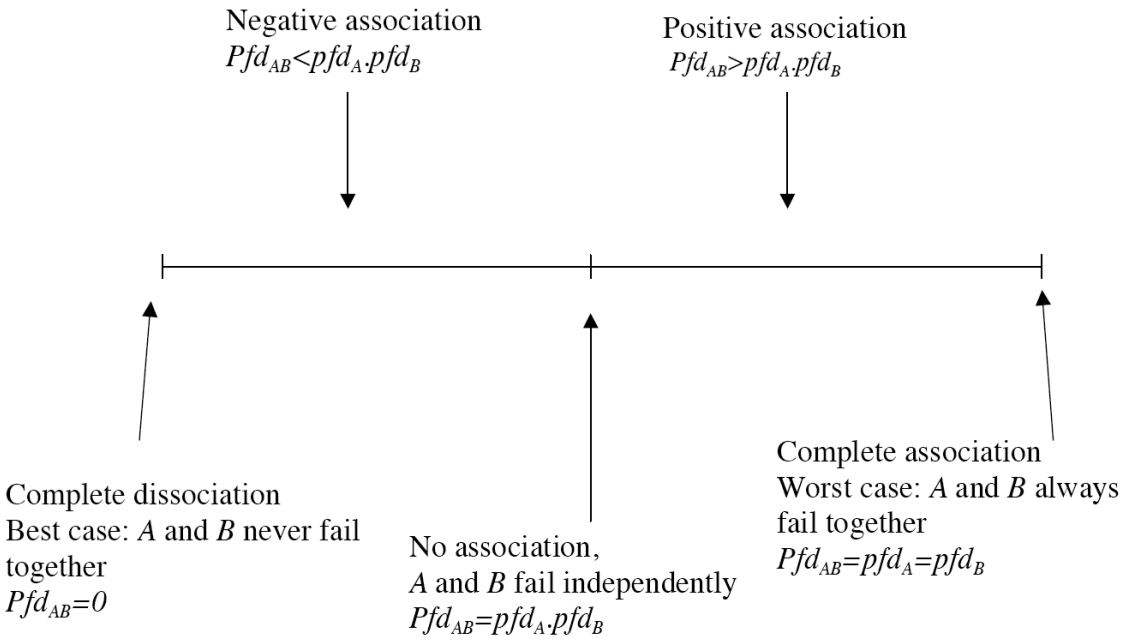

Figure 1 The spectrum of possible association between channel failures of a 1-out-of-2 diverse system.

For a particular pair of channels, there will be a point on the spectrum, x, that represents the degree of association between failures of that pair. We might express this numerically, for example, as the ratio $pfd_{AB}/pfd_A.pfd_B$ in which case the "spectrum" becomes the interval $(0,\ 1/\max(pfd_A, pfd_B))$.

The important point is that there is uncertainty about the value of x: an assessor could not be certain that x took a particular point value on the spectrum. It is appropriate, therefore, to treat x as a random variable, and an assessor can be expected to have

some prior beliefs about it (such beliefs might be based, for example, on knowledge of how the two channels were developed). As is often the case, it seems unlikely that an assessor would be able to state a complete prior *distribution* for x. We propose to examine the case where the assessor can tell us a single point on this distribution, specifically his probability that x is not greater than 1. This is his confidence that the association of channel failures is not positive in Figure 1 (that pfd_{AB} is not greater than $pfd_A.pfd_B$), i.e. channel failures are *not worse than independent* (NWTI).

Notice that in this case the assessor is expressing his belief about x as a probability associated with an *interval* on the spectrum of dependence. This seems more reasonable than associating a probability with a *point* on the spectrum – in particular with the "independence" point, x=1. We might be prepared to regard as reasonable an assessor's claim of the kind "I am 90% sure that x is no greater than 1", but not a claim such as "I am 90% sure that x=1". More formally, we assume that the distribution representing his belief about x is absolutely continuous, and thus has zero probability mass at a point.

2 Model based on an assessor's level of confidence that "channel failures are not worse than independent"

The basic idea here is similar to that we proposed in (Bishop, Bloomfield et al. 2011). In that paper it was shown how to obtain a conservative *pfd* for a single channel based on an assessor's prior belief and some failure-free operational testing. It was assumed that the assessor was only able to state a single percentile for his prior belief about the system *pfd*. It was shown that, of all complete prior distributions that satisfied this percentile constraint, the most conservative is a simple 2-point distribution. More surprisingly, it was also shown that following seeing the execution of some demands without failure, another 2-point prior distribution (satisfying the assessor's percentile constraint) gives the most conservative posterior mean *pfd*. The interpretation of this is that the assessor can treat this value as a conservative bound for his true probability of failure on demand – for example in a wider safety case.

We begin, for simplicity, by assuming that the channel *pfd*s are known with certainty: pfd_A, pfd_B. This assumption can later be relaxed by using the results of (Bishop, Bloomfield et al. 2011) upon each channel. So the only uncertainty concerns the degree of association between the failures of channel *A* and channel *B*. The assessor's confidence that there is not positive association between channel failures (NWTI) is $1\text{-}d_{AB}(0)$: so $d_{AB}(0)$ can be thought of as the assessor's doubt that the channel failures are independent *or better*[2].

That is, the assessor's prior distribution, $f(p)$, for the *system probability of failure*, PFD_{AB}[3], has a percentile:

$$P(PFD_{AB} \le pfd_A.pfd_B) = 1 - d_{AB}(0) \quad (1)$$

---

2 The notation here anticipates the more general one we require later in the paper, when we show how such confidence changes as a result of seeing N failure-free demands. Here N=0.

3 We shall use, as far as possible, upper case letters to indicate random variables, and lower case letters to represent their realisations.

We also assume initially, without loss of generality, that $pfd_A \leq pfd_B$. Since we know that the probability of failure of a 1-out-of-2 system cannot be worse than the best channel *pfd*, we have:

$$P(PFD_{AB} \leq pfd_A) = 1 \tag{2}$$

We thus have two percentiles of the assessor's prior distribution, $f(p)$, for the system's probability of failure on demand, PFD_{AB}. In general, there will be an infinite number of potential prior probability density functions, $f(p)$, that satisfy (1) and (2). It is easy to see that the most pessimistic of these is the 2-point distribution that has probability mass at $pfd_A.pfd_B$ (with probability $1–d_{AB}(0)$), and probability mass at pfd_A (with probability $d_{AB}(0)$).

Since the expert's probability that the system fails on a randomly selected demand is just the mean of his (in this case, prior) distribution of PFD_{AB}, we have:

$$\begin{aligned} &P(\text{system fails on randomly selected demand}) \\ &= E(PFD_{AB} \mid 0 \text{ failure-free tests observed}) = pfd_{AB}(0), \text{ say} \\ &= \int_0^1 p.f(p)\,dp \\ &\leq pfd_A.pfd_B(1 - d_{AB}(0)) + pfd_A.d_{AB}(0) \end{aligned} \tag{3}$$

This bound is the value that the assessor can treat as his true (prior) probability of failure of the system on a randomly selected demand, and be assured that it is a conservative (although attainable) number.

Of course, this prior bound may not be of practical value – it may be *very* conservative for reasonable values of $d_{AB}(0)$. However, as in the case of a single system (Bishop, Bloomfield et al. 2011), things become more interesting and useful when evidence is available of extensive failure-free working of the 1-out-of-2 system.

When N demands have been executed by the system, and no failures have been seen[4], the assessor's belief about PFD_{AB} changes from his prior distribution, $f(p)$, via Bayes' theorem. His posterior distribution is

$$f(p \mid N \text{ failure-free demands}) = \frac{(1-p)^N f(p)}{\int_0^1 (1-p)^N f(p)\,dp} \tag{4}$$

The assessor's posterior probability of failure on a randomly selected demand is the mean of this distribution:

$$\begin{aligned} &P(\text{System fails on randomly selected demand} \mid N \text{ failure-free demands}) \\ &= E(PFD_{AB} \mid N \text{ failure-free demands}) \\ &= \frac{\int_0^1 p(1-p)^N f(p)\,dp}{\int_0^1 (1-p)^N f(p)\,dp} \end{aligned} \tag{5}$$

---

4 Note we are assuming at this point that only *system* successes/failures are observed, and not the individual channel outcomes. That is, we are treating the system as a "black box".

The question now is which, of the infinite number of prior density functions f that satisfy (1) and (2), are the most pessimistic, i.e. maximize (5). In general there will be an infinite number of these. We can show that, once again, one of these is a 2-point distribution; i.e. this distribution has the same posterior expectation as the (many) most pessimistic priors. This distribution has probability mass concentrated at $pfd_A.pfd_B$, as before, and probability mass concentrated at a point $z_{AB}(N)$, where $z_{AB}(N)$ is the value of z that maximizes the posterior mean:

$$F_{AB}(z) = E(PFD_{AB} \mid N \text{ failure - free demands})$$

$$= \frac{pfd_A pfd_B (1 - pfd_A pfd_B)^N (1 - d_{AB}(0)) + z(1-z)^N d_{AB}(0)}{(1 - pfd_A pfd_B)^N (1 - d_{AB}(0)) + (1-z)^N d_{AB}(0)} \quad (6)$$

The conservative "true" system *pfd* is then the value (6) takes at its maximum, i.e.

$$E(PFD_{AB} \mid N \text{ failure - free demands}) \le F_{AB}(z_{AB}(N)) \quad (7)$$

$$= pfd_{AB}(N), \text{ say} \quad (8)$$

The assessor's doubt that the failures of the two channels are NWTI changes as he observes N failure-free demands, from his prior belief $d_{AB}(0)$ to posterior:

$$d_{AB}(N) = \frac{(1 - z_{AB}(N))^N d_{AB}(0)}{(1 - pfd_A pfd_B)^N (1 - d_{AB}(0)) + (1 - z_{AB}(N))^N d_{AB}(0)} \quad (9)$$

in an obvious notation. The conservative posterior distribution of the system *pfd*, (4), is a 2-point distribution with probability mass $(1\text{-}d_{AB}(N))$ at the "independence" point, $pfd_A.pfd_B$, and probability mass $d_{AB}(N)$ at $z_{AB}(N)$, which has the mean given by (7). For proof of these statements, see Appendix.

The reader should note that "conservative" here refers only to the *mean value* of the system *pfd*: from the infinite number of prior distributions that satisfy the assessor's expressed beliefs, (1) and (2), there is none that gives a larger posterior mean *pfd* than (8). The "conservative" posterior distribution here may not be conservative in other respects. For example, it has no probability mass to the right of the point $z_{AB}(N)$, which many assessors might regard as too optimistic.

3 Practical implications of the model: Examples

Consider the following simple illustrative example. Assume that $pfd_A = pfd_B = 10^{-3}$, the initial doubt, $d_{AB}(0) = 0.1$, and that we see 2000 failure-free demands, i.e. $N = 2000$. We have:

$pfd_{AB}(0) \sim 1\ 10^{-4}$, from (3)

$pfd_{AB}(2000) = 2.063\ 10^{-5}$

$d_{AB}(2000) = 0.0365$

$z_{AB}(2000) = 0.00054$

What has happened here – are these results useful? Clearly, the assessor's conservative "true" *prior* system probability of failure on demand, at approximately 10^{-4}, is not a very useful improvement on the single channel *pfd*s. However, the conservative "true" *posterior* system probability of failure on demand, $pfd_{AB}(2000)$, is almost two orders of magnitude better than the crude bound 10^{-3} (the probability of

failure of the best channel). It is also considerably better than the "black-box" 99% confidence bound 2.65 10^{-3}, which is obtained by treating the system as a single black box about which nothing is known except that it has survived 2000 demands without failure (Littlewood and Wright 1997).

Furthermore, confidence in "no worse than independence" of *A*, *B* channel failures has increased to over 0.96 from the assessor's original 0.90. The right hand point, z, of the conservative 2-point distribution has moved to the left, closer to the "independence" point $pfd_A.pfd_B$.

In general, the final (conservative) claim for system *pfd* depends upon: the channel probabilities of failure on demand, pfd_A, pfd_B; the prior doubt, $d_{AB}(0)$, about the channel failures being no worse than independent; and on the number of failure-free system demands, N, that have been seen. One way the results of Section 2 could be used would be to see whether any points in this (pfd_A, pfd_B, $d_{AB}(0)$, N) space seem feasible. For example, given the values of pfd_A, pfd_B, $d_{AB}(0)$, we could see how many test cases need to be executed (and show no failures) to obtain a particular value of $pfd_{AB}(N)$: in many cases, such as reactor protection systems, the cost of generating test cases may be high so that a large N may be infeasible. Alternatively, given the values of pfd_A, pfd_B, N (where here N is regarded as the size of the largest practically feasible test set), we could calculate the required $d_{AB}(0)$ and ask whether such a belief could be trusted (for example, supported by evidence about the diversity-seeking decisions (Littlewood and Strigini 2000; Wood, Belles et al. 2010) taken during system design and build).

We shall illustrate the general approach via the real example of a safety-critical protection system discussed briefly in Section 1. The aim was to claim a *pfd* of 10^{-6} for the two-channel system, each channel of which is software-based. The individual channel *pfd*s are estimated *conservatively* at 10^{-4} and 10^{-2} respectively. The system claim, 10^{-6}, is then obtained by multiplying these two channel claims. In our private discussions with safety engineers and assessors, we understood that the reasoning here is that "modest" dependence between channel failures will be more than countered by the conservatism of the individual channel *pfd* claims: the claim of 10^{-6} for the system *pfd* will then be conservative.

We have already expressed our scepticism about such a trade-off. We now sketch out how the system claim of 10^{-6} might be supported by the kind of reasoning of Section 2. In particular, we show how many failure-free demands of the system need to be observed to support the claim for different levels of doubt about "no worse than independence", and different degrees of conservatism in the channel *pfd* claims.

Table 1 shows the results of our analysis when the assessor believes the channel *pfd*s are no worse than 10^{-4} and 10^{-3} respectively. For the three different values of an assessor's doubt about NWTI, 0.10, 0.05, 0.01 respectively, the table shows the value of N for which the system *pfd* claim 10^{-6} can be supported. Thus, the number in bold for $F_{AB}(z_{AB}(N))$ in the second row of the table corresponds to N=43667, the smallest number of failure-free demands that allow a system *pfd* claim of better than 10^{-6} when the initial doubt is 0.10. These results are somewhat unforgiving: for the two most modest values of the doubt, the numbers of failure-free demands needed are rather high. This amount of operational testing may not be feasible for some applications. For example, it is an order of magnitude greater than what was feasible twenty years ago in the case of the Sizewell B PPS software (May, Hughes et al. 1995). However,

there have been significant advances in computing speeds in the past twenty years, and much larger simulations are now possible: for example, 50,000 test cases may be generated as part of the assessment of the C&I functions of the UK's proposed EPR (HSE 2011).

For the smallest doubt, 0.01 represented by the last two lines of the table, on the other hand, only 1055 failure-free demands are required, and this *does* seem sufficiently modest to be feasible in many cases.

$d_{AB}(0)$	N	$F_{AB}(z_{AB}(N))$	$d_{AB}(N)$	$z_{AB}(N)$
0.10	43666	1.000002e-06	0.03795288	2.381367e-05
0.10	43667	**9.999814e-07**	0.03795288	2.381313e-05
0.05	21108	1.000002e-06	0.01866105	4.832892e-05
0.05	21109	**9.999595e-07**	0.01866105	4.832663e-05
0.01	1054	1.000054e-06	0.009009550	1e-04
0.01	1055	**9.999650e-07**	0.009008659	1e-04

Table 1 Required values of N (number of failure-free demands) to support a system *pfd* claim of 10^{-6} for different values of initial assessor doubt $d_{AB}(0)$. Here $pfd_A=10^{-4}$, $pfd_B=10^{-3}$.

Table 2 shows the results for a similar calculation when the channel *pfd*s are no worse than $10^{-4.5}$ and $10^{-2.5}$ respectively. As in Table 1, these numbers are chosen so that their product – the "independence" case – is 10^{-7}. The final row of the table shows that, in this case, no failure-free demands are needed to make the conservative claim that the system *pfd* is better than 10^{-6}: this claim can be made simply from the prior beliefs about the channel *pfd*s when the doubt about NWTI is 0.01.

$d_{AB}(0)$	N	$F_{AB}(z_{AB}(N))$	$d_{AB}(N)$	$z_{AB}(N)$
0.1	43666	1.000002e-06	0.03795288	2.381367e-05
0.1	43667	**9.999814e-07**	0.03795288	2.381313e-05
0.05	18483	1.000018e-06	0.02855136	3.162278e-05
0.05	18484	**9.999907e-07**	0.02855049	3.162278e-05
0.01	0	**4.152278e-07**	0.01	3.162278e-05

Table 2 As Table 1, except $pfd_A=10^{-4.5}$, $pfd_B=10^{-2.5}$.

Finally, in Table 3 the results are shown for a calculation in which the channel *pfd*s are no worse than 10^{-5} and 10^{-2}. Again the product – the "independence" case – has been chosen to be 10^{-7}.

$d_{AB}(0)$	N	$F_{AB}(z_{AB}(N))$	$d_{AB}(N)$	$z_{AB}(N)$
0.1	10642	1.000000e-06	0.09909913	1e-05
0.1	10643	**9.999914e-07**	0.09909824	1e-05
0.05	0	**5.95e-07**	0.05	1e-05
0.01	0	**1.99e-07**	0.01	1e-05

Table 3 As Table 1, except $pfd_A=10^{-5}$, $pfd_B=10^{-2}$.

In this case the required conservative system *pfd* claim of no worse than 10^{-6} can be made for values of $d_{AB}(0)$ of 0.01 and 0.05 *a priori*, i.e. without seeing any failure-free working. Even when the doubt is 0.1, the required number of failure-free system demands is only 10642, which is more modest than the numbers required for the examples of Tables 1 and 2.

The results of Table 3 are less unforgiving than those of the other two tables. This seems to be because the channels are more asymmetric: channel *A* is much more reliable than channel *B*, and indeed pfd_A is only a single order of magnitude short of the overall *system* goal of 10^{-6}. Since the system *pfd* cannot be worse than the best channel *pfd*, quite modest confidence in NWTI means that the contribution from the second channel is sufficient to make the expected system *pfd* smaller than the required 10^{-6}.

These numbers are, of course, merely illustrative. They are intended to give the reader some feel for the trade-offs that are likely between "independence doubt", channel *pfd*s, and extensiveness of failure-free testing.

All the results above are obtained numerically: there is no closed form expression for $F_{AB}(z_{AB}(N))$. An alternative approach to the one above allows exact closed form results. It involves a kind of backward reasoning, in which an assessor – say a regulator – begins with a prior subjective doubt, say *D*, about NWTI, based on his review of the diversity-seeking practices adopted during the system development. That is

$$P(PFD_{AB} > pfd_A \times pfd_B) \le D \tag{10}$$

We assume that the system *pfd* requirement is P_{AB}, arising from the wider safety case, i.e.

$$E(PFD_{AB} \mid N \text{ failure-free demands}) \le P_{AB} \tag{11}$$

Then it can be shown that the upper bound on the prior doubt about NWTI required to satisfy (11) is

$$d_{req} = \frac{1}{1 + \dfrac{z_{AB}(N) - P_{AB}}{P_{AB} - pfd_A \times pfd_B} \times \left(\dfrac{1 - z_{AB}(N)}{1 - pfd_A \times pfd_B} \right)^N} \tag{12}$$

where

$$z_{AB}(N) = \min(pfd_A, pfd_B, z_m) \tag{13}$$

and

$$z_m = 1 - \left(1 - \frac{1}{N+1}\right)(1 - P_{AB})$$

Here $z_{AB}(N)$ is the upper point of support of the 2-point distribution that is the most pessimistic prior (the other point of support being $pfd_A \times pfd_B$), as before. This upper point of support will be z_m when this is smaller than each of the channel *pfd*s. For any given channel *pfd*s, this will happen when N is large enough, specifically when $N>N_C$, where

$$N_C = \frac{1 - \min(pfd_A, pfd_B)}{\min(pfd_A, pfd_B) - P_{AB}} \tag{14}$$

In that case

$$d_{req} = \frac{1}{1 + \frac{1 - P_{AB}}{P_{AB} - pfd_A \times pfd_B} \times \left(\frac{1 - P_{AB}}{1 - pfd_A \times pfd_B}\right)^N \times \left(1 + \frac{1}{N}\right)^{-N} \times \frac{1}{N+1}} \tag{15}$$

For proofs, and details of closed form expressions for d_{req}, see Appendix.

All this might be used in a two-stage procedure as follows. An assessor, such as a regulator, having arrived at a probability D that represents his prior doubt about NWTI, would compute d_{req} (based on the known values of pfd_A, pfd_B, N and P_{AB}) and compare this with D. If $d_{req}<D$ he would reject the claim P_{AB}, (11). If $d_{req}\geq D$ he would accept the claim.

Tables 4, 5, 6 show examples, continuing the example introduced in Section 1. As before, the system *pfd* requirement is 10^{-6}, and the channel *pfd*s have been chosen in each row of each table to give a product of 10^{-7} in the spirit of "conservatism about channel claims". In the first column of each table, the successive values of pfd_A from the top are 10^{-2}, $10^{2.2}$, $10^{-2.4}$, … , 10^{-5}; the values of pfd_B are in the same range, but starting from the bottom, so that for each row $pfd_A \times pfd_B = 10^{-7}$. The three tables differ in the number, N, of failure-free demands observed.

The tables show clearly the way in which asymmetry in the channel *pfd*s aids the assessment: the more asymmetric these are, all things being equal, the greater the prior doubt about NWTI can be, whilst still allowing the claim about the system *pfd*. Thus in Table 4, the greatest doubt that can be allowed occurs when $pfd_A=10^{-2}$, $pfd_B=10^{-5}$ (or vice-versa). Similar results apply in Tables 5 and 6 although, for these larger values of N, the differences between the largest allowable doubt and the smallest, over the range of values of the channel *pfd*s, is less pronounced.

Notice that the value of d_{req} for large values of N, (15), depends on the channel probabilities of failure on demand, pfd_A and pfd_B, only via their product. The extent to which this product is smaller than P_{AB} can be thought of as representing the degree of conservatism in the system *pfd* claim, compared with an over-optimistic claim of *certain* independence of channel failures. This is similar to the informal reasoning we reported in Section 1, but in our more formal treatment, the system claim is *guaranteed* to be conservative (for the assessor's particular level of doubt about NWTI).

In Table 1, the central rows all have d_{req}=0.024. That is because for these values of the channel *pfd*s, *N*=10000 is sufficiently large to satisfy (14) – i.e. z_m is smaller than each channel *pfd* – and so in (15) d_{req} depends on the individual channel *pfd*s only via their product, which is 10^{-7} in each row. This effect is even more pronounced in Tables 5 and 6, in which *N* is larger.

pfd_A	pfd_B	P_{AB}	N	d_{req}
0.01	1e-05	1e-06	10000	0.099
0.006309573	1.584893e-05	1e-06	10000	0.066
0.003981072	2.511886e-05	1e-06	10000	0.046
0.002511886	3.981072e-05	1e-06	10000	0.033
0.001584893	6.309573e-05	1e-06	10000	0.026
0.001	1e-04	1e-06	10000	0.024
0.0006309573	0.0001584893	1e-06	10000	0.024
0.0003981072	0.0002511886	1e-06	10000	0.024
0.0002511886	0.0003981072	1e-06	10000	0.024
0.0001584893	0.0006309573	1e-06	10000	0.024
1e-04	0.001	1e-06	10000	0.024
6.309573e-05	0.001584893	1e-06	10000	0.026
3.981072e-05	0.002511886	1e-06	10000	0.033
2.511886e-05	0.003981072	1e-06	10000	0.046
1.584893e-05	0.006309573	1e-06	10000	0.066
1e-05	0.01	1e-06	10000	0.099

Table 4: Values of d_{req}, i.e. prior doubt about NWTI, that must not be exceeded in order to support a system *pfd* claim of 10^{-6} after seeing 10,000 failure-free demands, for different values of the channel *pfd*s (where, in each case, the product of the channel *pfd*s is 10^{-7}). Here z_m=1.0099e-04.

pfd_A	pfd_B	P_{AB}	N	d_{req}
0.01	1e-05	1e-06	50000	0.141
0.006309573	1.584893e-05	1e-06	50000	0.118
0.003981072	2.511886e-05	1e-06	50000	0.113
0.002511886	3.981072e-05	1e-06	50000	0.113
0.001584893	6.309573e-05	1e-06	50000	0.113
0.001	1e-04	1e-06	50000	0.113
0.0006309573	0.0001584893	1e-06	50000	0.113
0.0003981072	0.0002511886	1e-06	50000	0.113
0.0002511886	0.0003981072	1e-06	50000	0.113
0.0001584893	0.0006309573	1e-06	50000	0.113
1e-04	0.001	1e-06	50000	0.113
6.309573e-05	0.001584893	1e-06	50000	0.113
3.981072e-05	0.002511886	1e-06	50000	0.113
2.511886e-05	0.003981072	1e-06	50000	0.113
1.584893e-05	0.006309573	1e-06	50000	0.118
1e-05	0.01	1e-06	50000	0.141

Table 5: As Table 4, but *N*=50,000, and z_m=2.1e-05.

pfd_A	pfd_B	P_{AB}	N	d_{req}
0.01	1e-05	1e-06	1e+05	0.212
0.006309573	1.584893e-05	1e-06	1e+05	0.211
0.003981072	2.511886e-05	1e-06	1e+05	0.211
0.002511886	3.981072e-05	1e-06	1e+05	0.211
0.001584893	6.309573e-05	1e-06	1e+05	0.211
0.001	1e-04	1e-06	1e+05	0.211
0.0006309573	0.0001584893	1e-06	1e+05	0.211
0.0003981072	0.0002511886	1e-06	1e+05	0.211
0.0002511886	0.0003981072	1e-06	1e+05	0.211
0.0001584893	0.0006309573	1e-06	1e+05	0.211
1e-04	0.001	1e-06	1e+05	0.211
6.309573e-05	0.001584893	1e-06	1e+05	0.211
3.981072e-05	0.002511886	1e-06	1e+05	0.211
2.511886e-05	0.003981072	1e-06	1e+05	0.211
1.584893e-05	0.006309573	1e-06	1e+05	0.211
1e-05	0.01	1e-06	1e+05	0.212

Table 6: As Table 4, but *N*=100,000 and z_m=1.1e-05.

There is an obvious interplay here between channel *pfd* asymmetry and the size of *N*: high asymmetry essentially means that one of the *pfd*s is very small, and N_C in (14) is thus large. So, in Table 4, where *N* is of quite modest size, the "best" – i.e. the largest – acceptable doubt is a factor of four greater than that where the channel *pfd*s are approximately equal in size.

In Table 6, in contrast, *N* is sufficiently large that for almost all values of the channel *pfd*s it is greater than N_C and so d_{req} takes the same value in almost all cases. That is, channel asymmetry cannot be exploited here to increase allowable doubt in NWTI. Or, putting it more positively, for such large *N* there is no need to have one channel very much more reliable than the other to gain benefit in the size of d_{req}.

Tables 7, 8, 9 show similar results in a case where there is greater conservatism in the channel *pfd* claims: here the product is 10^{-8} in contrast to the 10^{-7} of the previous tables.

pfd_A	pfd_B	P_{AB}	N	d_{req}
1e-03	1e-05	1e-06	10000	0.10838
6.3096e-04	1.5849e-05	1e-06	10000	0.072454
3.9811e-04	2.5119e-05	1e-06	10000	0.050118
2.5119e-04	3.9811e-05	1e-06	10000	0.036589
1.5849e-04	6.3096e-05	1e-06	10000	0.02909
1e-04	1e-04	1e-06	10000	0.026462
6.3096e-05	1.5849e-04	1e-06	10000	0.02909
3.9811e-05	2.5119e-04	1e-06	10000	0.036589
2.5119e-05	3.9811e-04	1e-06	10000	0.050118
1.5849e-05	6.3096e-04	1e-06	10000	0.072454
1e-05	1e-03	1e-06	10000	0.10838

Table 7: Similar to Table 4, but with the channel *pfd*s in each row having a product of 10^{-8}, i.e. there is greater conservatism in the channel *pfd* claims.

pfd_A	pfd_B	P_{AB}	N	d_{req}
1e-03	1e-05	1e-06	50000	0.15345
6.3096e-04	1.5849e-05	1e-06	50000	0.12831
3.9811e-04	2.5119e-05	1e-06	50000	0.12387
2.5119e-04	3.9811e-05	1e-06	50000	0.12387
1.5849e-04	6.3096e-05	1e-06	50000	0.12387
1e-04	1e-04	1e-06	50000	0.12387
6.3096e-05	1.5849e-04	1e-06	50000	0.12387
3.9811e-05	2.5119e-04	1e-06	50000	0.12387
2.5119e-05	3.9811e-04	1e-06	50000	0.12387
1.5849e-05	6.3096e-04	1e-06	50000	0.12831
1e-05	1e-03	1e-06	50000	0.15345

Table 8: As Table 7, but with *N*=50000.

pfd_A	pfd_B	P_{AB}	N	d_{req}
1e-03	1e-05	1e-06	100000	0.23001
6.3096e-04	1.5849e-05	1e-06	100000	0.22906
3.9811e-04	2.5119e-05	1e-06	100000	0.22906
2.5119e-04	3.9811e-05	1e-06	100000	0.22906
1.5849e-04	6.3096e-05	1e-06	100000	0.22906
1e-04	1e-04	1e-06	100000	0.22906
6.3096e-05	1.5849e-04	1e-06	100000	0.22906
3.9811e-05	2.5119e-04	1e-06	100000	0.22906
2.5119e-05	3.9811e-04	1e-06	100000	0.22906
1.5849e-05	6.3096e-04	1e-06	100000	0.22906
1e-05	1e-03	1e-06	100000	0.23001

Table 9: As Table 7, but with *N*=100,000.

From a practical viewpoint an important question is how best to build (and test) a system so that the resulting d_{req} is larger than D. It seems that asymmetry of the channel *pfd*s might be helpful here. However, since this essentially means that one of the channel *pfd*s needs to be close to the required *system pfd*, it may not be a practical proposition in cases where very high system reliability is needed. In fact, such asymmetry goes against the spirit of this kind of fault tolerance, which is to build highly reliable systems from channels of only modest reliability.

The other factors affecting d_{req} are the conservatism of the system *pfd* claim (i.e. how much it differs from the too-optimistic simple product of the channel *pfd*s), and the number of (failure-free) test cases observed. A comparison between Tables 4-6 and Tables 7-9 indicates the advantage, in terms of larger d_{req}, when there is greater conservatism (i.e. when the product of the channel *pfd*s is 10^{-8} rather than 10^{-7}).

Table 10 summarises the interplay between the product $pfd_A \times pfd_B$, N and d_{req}: different levels of conservatism ($pfd_A \times pfd_B = 10^{-7}, 10^{-7.5}, 10^{-8}$), to support the same system *pfd* claim of 10^{-6}, are shown against their corresponding values of the doubt in NWTI needed for different values of N.

$pfd_A \times pfd_B$	P_{AB}	N	d_{req}
1e-7.0	1e-6	10000	0.099
1e-7.5	1e-6	10000	0.106
1e-8.0	1e-6	10000	0.108
1e-7.0	1e-6	50000	0.141
1e-7.5	1e-6	50000	0.15
1e-8.0	1e-6	50000	0.153
1e-7.0	1e-6	100000	0.212
1e-7.5	1e-6	100000	0.225
1e-8.0	1e-6	100000	0.230

Table 10: Upper bounds on the prior doubt about NWTI required to support a system *pfd* claim of 10^{-6}, for three different values of $pfd_A.pfd_B$, for three different values of N.

Readers might well ask at this stage whether this new approach could be used in practice for the assessment of real systems, bearing in mind that for safety critical applications the system *pfd* requirement may be a stringent one. The numerical value of P_{AB} used in our examples here is, as we have said, one that we know to be the requirement for a real critical system. The issue then is whether the numbers in the tables above are plausible ones to be part of a safety assessment to support a claim of this magnitude. That is, for a particular instance (i.e. a row of one of the Tables 4-9): Are the required pfd_A, pfd_B achievable (and assessable)? Is the number of tests, N,

feasible? And, most importantly, is it believable that the assessor's doubt D in NWTI is smaller than that required, d_{req}?

Of course, such questions can really only be answered when there is specific evidence available about a particular system. However, we believe – somewhat tentatively – that it is reasonable to answer in the affirmative in some of the cases above. Take Table 9. Here there is considerable conservatism in the channel claims (product equals 10^{-8} versus a system claim of 10^{-6}), so that an assessor's doubt about NWTI can be as high as 23% and still allow him to accept the system claim. He can do this without appealing to channel asymmetry (i.e. an implausibly strong claim for one of the channels), because the middle row of the table shows that claims of 10^{-4} for each channel will be sufficient. Such claims seem relatively modest for channels that have been built to safety-critical standards: for example, they could be supported by feasible amounts of operational testing. The number of *system* test cases (100000) that need to be generated is large, of course. Whether this is feasible will depend on particular circumstances, but we note that for a real protection system it is proposed to generate 50000 test cases (HSE 2011). Notice, however, that even 100000 test cases is more than an order of magnitude fewer than would be needed to support the system *pfd* claim of 10^{-6} directly from a "black-box" test (Littlewood and Wright 1997).

We think the greatest difficulty in using a model like this will lie in an assessor arriving at a believable numerical value for D. We discuss this issue briefly in the next section.

4 Discussion

We have presented a new way of reasoning about the reliability of a 2-channel, 1-out-of-2 software-based system that overcomes some of the objections that can be made about an earlier approach to the problem. This approach can be characterized as follows: "We realize that an assumption of independence between failures (and thus a claim for system *pfd* that is the simple product of channel *pfd*s) may be too optimistic, but we have compensated for that by making only very conservative claims for the channel *pfd*s. The system *pfd* claim will thus be conservative."

We believe the work reported here captures the spirit of this informal reasoning, but does so in a way that is more rigorous and thus believable. It gives a rigorous meaning to notions of "conservative", and allows proper trade-offs to be made between the different model parameters (prior doubt about NWTI, number of failure-free tests observed, product of channel *pfd*s, system *pfd* claim). However, the attentive reader will have noticed that this new approach brings its own problems and some difficulties that need further thought.

In the first place, we have assumed in the development of Section 2 that the channel probabilities of failure on demand, pfd_A and pfd_B, are known. In practice, of course, these probabilities will not be known with certainty. One way forward would be to use the ideas in (Bishop, Bloomfield et al. 2011), where it was shown how to obtain a conservative bound for the posterior mean of the *pfd* of a single system based on an assessor's prior belief and the observation of some failure-free demands in statistically representative operational testing. Such bounds could be used by an assessor as if they were "true" *pfd*s, in the knowledge that they will be conservative.

The testing of the different channels to obtain these *pfd*s would be carried out before the system testing required for the results of Sections 2 and 3. In the current model there is no further "learning" about these channel *pfd*s from the *system* testing. That is, the likelihood function used in the Bayesian updating in Section 2 does not allow any updating of the assessor's knowledge of his beliefs about the channel *pfd*s. The evidence from the system testing is just that there have been no system failures in N tests, but the assessor does not know whether there have been individual channel failures. Informally, the Bayesian updating in Section 2 concerns only the channel failure dependence via the evolution of $d_{AB}(N)$ and $z_{AB}(N)$ as N increases, but not any evolution of the channel reliabilities.

This may be realistic in some cases: for example, in the case of shut-down systems, when a preferred channel correctly causes shut-down, the other channel may not be invoked and so it may not be known if it would have failed to shut down on that demand. However, in many cases this view will be too restrictive, and the *system* tests will also give information about *channel* outcomes. In such cases it would be useful to extend the model to be able to take account of this information: this is an issue we plan to address in future work. In fact such an extended model may also allow greater confidence to be gained in NWTI: informally, seeing some single channel failures, but no system failures, may give greater confidence in the efficacy of the fault tolerance mechanism (albeit less confidence in the reliabilities of the channels).

Another difficulty concerns the assessor's prior doubt about no-worse-than-independence of channel failures. Is it reasonable to expect an assessor to be able to state a numeric value for D, and for this number to be genuinely meaningful, rather than simply an uninformed guess? Interestingly, in private discussions with safety engineers and regulators familiar with these kinds of multi-channel systems we have found a willingness to express numerically their confidence about independence itself: e.g. "I am 90% confident that failures of these channels will be independent". In supporting such claims the experts usually appeal to their detailed knowledge of the architectures of the target systems, and on how they were built.

As we have argued here, claims about independence itself – a point on the dependence spectrum – do not seem realistic, and a better way to proceed is to make claims that are framed in terms of *intervals*, such as our "no worse than independence". It is interesting to ask whether the same experts would be able to support *these* claims using evidence from system architectures, and details of the design and build processes. Using such evidence to support probabilistic measures of doubt, as is required here, may not be easy.

Empirical evidence of the levels of dependence between diverse software-based channels is *very* thin on the ground. A single data point comes from the multi-version experiment conducted by Knight and Leveson (Knight and Leveson 1986). There the null hypothesis of NWTI was not rejected for 139 out of 162 pairs of versions. That is, the estimated doubt in NWTI for a randomly selected pair was 0.142 (Povyakalo and Littlewood 2010) which compares favourably with the 0.23 doubt in the discussion of the previous section. However, it has to be said that the problem addressed here was not comparable to a real safety-critical application, such as a protection system, and the versions were not developed under the kind of conditions that might be expected of such applications.

Ideally, we would like to have empirical evidence of the channel dependencies achieved in *real* systems. We are not aware of such evidence being available currently, in spite of several well-known multi-channel software-based systems having received extensive operational exposure. What is needed is that channel "vote-outs" be recorded in those situations where there is no *system* failure (as well as when there *is* system failure, of course). We are not aware that this is done as a matter of course in any existing systems, and have seen no published data of this kind.

Even if such data were available, across many disparate safety-critical systems in operation, there would be difficulties in an assessor using them to make a judgment about his confidence in NWTI for a particular *novel* system, since this new one may differ in significant ways from the previous ones.

However, in some industries, there may be considerable experience in building "similar" systems in the past: e.g. protection systems. Let us assume that these earlier systems have been successful, in the sense that they were accepted as sufficiently safe to be deployed, and these judgements were not overruled by operational experience. An assessor could retrospectively compute d_{req} for each of these previous systems. It might then be conservative for him to use, say, the smallest of these numbers as his prior doubt, D, for a *novel* system. Such a choice could be regarded as ensuring that the procedure for assuring the safety of a new system was no worse – i.e. no less stringent – than that adopted historically. Such an approach seems to fit well with the UK principle of ALARP (As Low As Reasonably Practicable).

Acknowledgements

We would like to thank several colleagues for extensive discussions about the issues dealt with in this paper: Bob Jennings and Bob Yates from Office of Nuclear Regulation; Silke Kuball from EDF; Peter Bishop, Lorenzo Strigini and Robin Bloomfield from CSR.

Support for the work reported here came from:

- the UnCoDe project, funded by the Leverhulme Trust;
- the DISPO project, funded under the CINIF Nuclear Research Programme by EDF Energy Limited, Nuclear Decommissioning Authority (Sellafield Ltd, Magnox Ltd), AWE plc and Urenco UK Ltd ("the Parties"). The views expressed in this Report are those of the author(s) and do not necessarily represent the views of the members of the Parties. The Parties do not accept liability for any damage or loss incurred as a result of the information contained in this Report.

References

Bishop, P., R. Bloomfield, et al. (2011). "Towards a formalism for conservative claims about the dependability of software-based systems." IEEE Trans Software Engineering **37**(5): 708-717.

Eckhardt, D. E., A. K. Caglayan, et al. (1991). "An experimental evaluation of software redundancy as a strategy for improving reliability." IEEE Trans Software Eng **17**(7): 692-702.

Eckhardt, D. E. and L. D. Lee (1985). "A Theoretical Basis of Multiversion Software Subject to Coincident Errors." IEEE Trans. on Software Engineering **11**: 1511-1517.

HSE (2011). Step 4 Control and Instrumentation Assessment of the EDF and AREVA UK EPR Reactor. Bootle, Health and Safety Executive, Office for Nuclear Regulation.

Knight, J. C. and N. G. Leveson (1986). An Empirical Study of Failure Probabilities in Multi-version Software. Proc. 16th Int. Symp. on Fault-Tolerant Computing (FTCS-16), Vienna, Austria.

Knight, J. C. and N. G. Leveson (1986). "Experimental evaluation of the assumption of independence in multiversion software." IEEE Trans Software Engineering **12**(1): 96-109.

Littlewood, B. and D. R. Miller (1989). "Conceptual Modelling of Coincident Failures in Multi-Version Software." IEEE Trans on Software Engineering **15**(12): 1596-1614.

Littlewood, B., P. Popov, et al. (2002). "Modelling software design diversity - a review." ACM Computing Surveys **33**(2): 177-208.

Littlewood, B. and L. Strigini (2000). A discussion of practices for enhancing diversity in software designs. London, Centre for Software Reliability, City University**:** 58.

Littlewood, B. and D. Wright (1997). "Some conservative stopping rules for the operational testing of safety-critical software." IEEE Trans Software Engineering **23**(11): 673-683.

May, J., G. Hughes, et al. (1995). "Reliability estimation from appropriate testing of plant protection software." Software Engineering Journal **10**(6): 206-218.

Povyakalo, A. and B. Littlewood (2010). Probably independent random variables and conservative bounds for expectation of their product. London, Center for Software Reliability, City University.

Wood, R. T., R. Belles, et al. (2010). Diversity Strategies for Nuclear Power Plant Instrumentation and Control Systems. Washington, DC, US Nuclear Regulatory Commission.